生命システム解明の基盤データベース構築

●金久 實 [1,2)]　◆服部 正泰 [1)]　片山 俊明 [2)]
1) 京都大学化学研究所バイオインフォマティクスセンター　2) 東京大学医科学研究所ヒトゲノム解析センター

＜研究の目的と進め方＞

本研究は、遺伝子･分子レベルの網羅的な解析から、細胞･個体･生態系レベルでの生命システムの全体像を明らかにすることを目指し、新しい情報技術の開発とともに、新しいタイプの基盤データベースを構築することを目的とする。細胞・個体レベルでの生命システム情報は、これまでのKEGGにおいてすでにデータベース化が行われているので、本研究では生物種間相互作用や環境との相互作用といったより高次レベルの生命システム情報をゲノムの情報と統合し、医療や産業をはじめ、ゲノム情報の有効利用へつなぐ基盤データベースを構築する。同時に支援班との協力の下に、様々な利用・解析ツールを開発して、特定ゲノム4領域との間でフィードバック連携をはかり、これら領域の研究推進及び成果の統合化に寄与する。

＜2007年度の研究の当初計画＞

文献データやなまの実験データを蓄積した従来型のデータベースに対し、本研究のデータベースはこれらのデータから得られる「知識」を蓄積する。その知識をもとに新しい研究を推進することが可能となるように、本データベースの様々な利用ツールを開発する。本計画の内容は以下の通りである。

・生命システムを構成する部品の情報として、ゲノム情報から直接的に規定される遺伝子とタンパク質はKEGG GENESデータベースに、それ以外のケミカル情報に関連した分子はKEGG LIGANDデータベースに蓄積する。また部品間のネットワーク情報はKEGG PATHWAYデータベースに蓄積する。
・GENESの情報はKEGG ORTHOLOGY (KO)として体系化し、これをもとにしたゲノムアノテーションシステムKAASをさらに高度化し、パスウェイマッピングだけでなくBRITEマッピングを充実させる。またメタゲノムのデータを解析する方法論の開発を行う。
・LIGANDの化学反応情報はRDMパターンとして体系化し、これをもとにしたケミカルアノテーションシステムとして既存のPATHCOMPとe-zymeを統合した反応予測システムの開発を行う。またゲノムの情報から糖鎖構造、脂質構造、ポリケチド・非リボソームペプチド構造などの化学構造を予測する方法論を精密化し、実用的なツール開発を行う。
・ゲノムネット（http://www.genome.jp/）のWebサービスを通して、KEGGをはじめとしたリソースをプログラムから呼び出し、カスタマイズして利用できるインターフェースKEGG APIの開発を継続して行う。
・基盤ゲノム領域及び他のゲノム領域に対して、ゲノムアノテーション、ケミカルアノテーションなど、本研究の成果を生かした支援を行う。また、KEGGの利用講習会やKEGG API入門コース等を開催して、その普及をはかる。
・高度の専門知識を効率的に集積するために、研究コミュニティと密接に連携し、その知識を集約する「コミュニティデータベース」の枠組みとして、CYORFデータベース等をさらに発展させる。

＜2007年度の成果＞

KEGG GENESデータベースには630生物種の285万遺伝子に関する情報が蓄積され、約11,000のKOオーソロググループに分類されている。昨年度と比較してGENESは52%、KOは15%程度の増加である。昨年度と一昨年度の比較ではGENESは49%、KOは16%の増加であったので、ほぼ同じような増加傾向にある。これまではすべての生物種に手作業でKOづけを行ってきたが、データ量の増加でこれは実質的に不可能になりつつあることから､KAAS自動アノテーションプログラム [3] の改良を行っている。とくにKEGGアノテータの知識を取り込んで、できるだけカバー率をあげ、精度をあげることに努めている。

KEGG LIGANDデータベースでは、ENZYMEデータベースを命名委員会のサイト (Trinity College Dublin) から毎週更新し、REACTIONデータと生物種の情報を付加する体制を作った。さらにREACTIONから作られるRPAIRのデータを全面的に見直し、反応に伴う化合物の構造変化パターンであるRDMパターンのアノテーションを行った。これを用いて反応予測プログラムe-zymeの新バージョンを提供した。現在はRDMパターン間の類似性解析と階層化を行っている。

KEGG PATHWAYデータベースでは、植物の二次代謝経路を重点的に整備した。また、既存の約120枚の代謝パスウェイマップを手作業でまとめたグローバルマップ [9] を構築し、KEGG Atlasと呼ぶ新機能として公開した。代謝のグローバルマップはSVGファイルで作られており、Google Mapsのようにズームして閲覧し、既存のマップやBRITE機能階層と統合して利用することができるソフトウェアもあわて開発した。グローバルマップは今後のKEGG利用の主流になるものと考えられ、実際Peer Borkのグループではすでに独自にソフトウェア開発を行い、メタゲノム解析等に利用している。

KEGGでは2種類のケミカルアノテーションに関する情報技術開発を行ってきた。1つはゲノムやトランスクリプトームの情報から生体内物質の化学構造を予測することで、すでに昨年度までに糖転移酵素のレパートリーから糖鎖構造のレパートリーを関連づける方法論の開発し実用化した。今年度は、ポリケチド・非リボソームペプチド [1] および不飽和脂肪酸 [8] で類似の方法論を開発した。もう1つのケミカルアノテーションは化合物（とくに生体外物質）の化学構造から生体システムとの相互作用を予測するもので、微生物による環境物質分解経路予測 [2] と内分泌撹乱物質であるかの判別 [7] を行った。

支援活動については、本特定領域の個別の研究グループに対してcDNAデータのアノテーション支援などを行い、また以下の公開行事を開催した。

(1) KEGGデータベース利用講習会

8月30日と31日に「第1回KEGGデータベース利用講習会」を京大化研バイオインフォマティクスセンターにおいて開催し、EGassembler, KAAS, GENIESを中心とした実習を行った。参加者は35名。また､11月15日と16日に「第2回KEGGデータベース利用講習会」を東大医科研ヒトゲノム解析センターにおいて開催し、医薬品や化合物を中心とした実習を行った。参加者は21名。いずれの回もWeb等で開催案内を出すと直ちに定員に達してしまうほどの盛況であった。

(2) UniProtデータベース利用講習会

10月2日に京大化研バイオインフォマティクスセンターで、4日と5日に東大医科研ヒトゲノム解析センターで「UniProtデータベース利用講習会」を開催した。これはKEGGとUniProtの連携の一貫として行ったものである。参加者は京都が10名。東京が延べ22名。

＜国内外での成果の位置づけ＞

KEGGは生命システムをコンピュータの中に再現した「生命システム情報統合データベース」で、ゲノム解読の最先端のリソースとして国際的に広く利用されている。NCBIやUniProtなど欧米の代表的データベースと協力関係にあり、相互にリンクづけがなされている。また、米国Consortium for Functional Glycomicsとの連携によるKEGG GLYCANや、IUPAC/IUBMB生化学命名委員会との連携によるKEGG ENZYMEをはじめ、様々なグループとも協力関係にある。生命システム解明に必要な情報をゲノム情報（KEGG GENES)、ケミカル情報（KEGG LIGAND)、システム情報（KEGG PATHWAY / KEGG BRITE）に分け、コンピュータ化が遅れているシステム情報とケミカル情報に重点を置いてデータベース化を行っていることが、国際的に評価されている。

＜達成できなかったこと、予想外の困難、その理由＞

Webのフィードバックシステムで寄せられる意見や、利用講習会での反響から、国内でも個人レベルのユーザはKEGGを頻繁に利用し、またKEGGに対する期待も大きいようである。しかしながら基盤ゲノム領域あるいはゲノム4領域研究参加者によるKEGGの利用は未だ不十分である。大きなプロジェクトでは協力体制を組み直して、KEGGのリソースを取り入れる形にするのは時間的に困難なのかもしれない。本計画研究では、KAASやKEGG Atlasといった新しいツールを開発し、利用講習や普及活動も活発に行ってきた。我が国全体の情報インフラ整備という面では大きく貢献している。

＜今後の課題＞

当初は皆無であったパスウェイデータベースも今では多数存在し、KEGGが常に最先端であるためには、次のステップの研究を可能にするリソースを開拓し続ける必要がある。そのため、KEGGが多様化したことによる使い勝手の悪さも出てきており、今後の課題は使いやすいインターフェースの開発である。KEGGの利用は単にデータベースを閲覧･検索するだけでなく、ゲノム、トランスクリプトーム、メタボロームといった大量の分子情報をマッピングして、高次生命システム機能に関する手がかりを得ることである。マッピングの対象となるデータベースは当初はPATHWAYだけであったが、現在はBRITEさらにはMODULEも加わって3種類になった。これらを統合的に扱うのがKEGG Atlasである。様々なグローバルマップのデータベース化とともに、Google MapsやGoogle Earthのようなソフトウェアを開発していきたいと考えている。

＜成果公表リスト＞

論文

1.0704271733
Minowa, Y., Araki, M., and Kanehisa, M.; Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J. Mol. Biol. 368, 1500-1517 (2007).

2.0704271736
Oh, M., Yamada, T., Hattori, M., Goto, S., and Kanehisa, M.; Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. J. Chem. Inf. Model. 47, 1702-1712 (2007).

3.0704271738
Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A., and Kanehisa, M.; KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35, W182-W185 (2007).

4.0708091337
Fujita, M., Mihara, H., Goto, S., Esaki, N., and Kanehisa, M.; Mining prokaryotic genomes for unknown amino acids: a stop-codon-based approach. BMC Bioinformatics 8, 225 (2007).

5.0708091346
Itoh, M., Nacher, J.C., Kuma, K.I., Goto, S., and Kanehisa, M.; Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. Genome Biol. 8, R121 (2007).

6.0708091351
Limviphuvadh, V., Tanaka, S., Goto, S., Ueda, K., and Kanehisa, M.; The commonality of protein interaction networks determined in Neurodegenerative disorders (NDDs). Bioinformatics 23, 2129-2138 (2007).

7.0801200905
Kadowaki, T., Wheelock, C.E., Adachi, T., Kudo, T., Okamoto, S., Tanaka, N., Tonomura, K., Tsujimoto, G., Mamitsuka, H., Goto, S., and Kanehisa, M.; Identification of endocrine disruptor biodegradation by integration of structure-activity relationship with pathway analysis. Environ. Sci. Technol. 41, 7997-8003 (2007).

8.0801200915
Hashimoto, K., Yoshizawa, A.C., Okuda, S., Kuma, K., Goto, S., and Kanehisa, M.; The repertoire of desaturases and elongases reveals fatty acid variations in 56 eukaryotic genomes. J. Lipid Res. 49, 183-191 (2008).

9.0801200923
Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y.; KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480-D484 (2008).